> Конспект > 4 урок > Стратификация

> Оглавление

- > Оглавление
- > Стратификация
  - > Как проводить стратификацию?
  - > Популяционное среднее
  - > Точечные оценки
  - > Реализация
- > Стратифицированное семплирование
- > Условное математическое ожидание
  - > Полные математическое ожидание и дисперсия
    - > Закон полного математического ожидания
- > Понижение дисперсии
  - > Межгрупповая и внутригрупповая дисперсия
    - > Дисперсия случайного семплирования
    - > Дисперсия стратифицированного семплирования
  - > Понижение дисперсии
  - > Преимущества стратифицированного семплирования
- > Постстратификация
  - > Проблемы случайного разбиения
  - > Дисперсия при постстратификации
- > Сравнение методов семплирования
  - > Дисперсии для различных методов
  - > Соотношения дисперсий
  - > Оценка пилота
- > Резюме
- > Материалы для самостоятельного изучения

> Стратификация

Ситуация: хотим изменить рекламу встроенных покупок, чтобы увеличить их продажи.

Мы уже определили метрики, размер пилотной и контрольной групп. Теперь хотим отнести к ним пользователей. Как это сделать?

Первое, что приходит на ум — случайно распределяем пользователей по группам и начинаем пилот.

Однако пользователи могут отличаться по различным характеристикам: полу, возрасту и т.д.

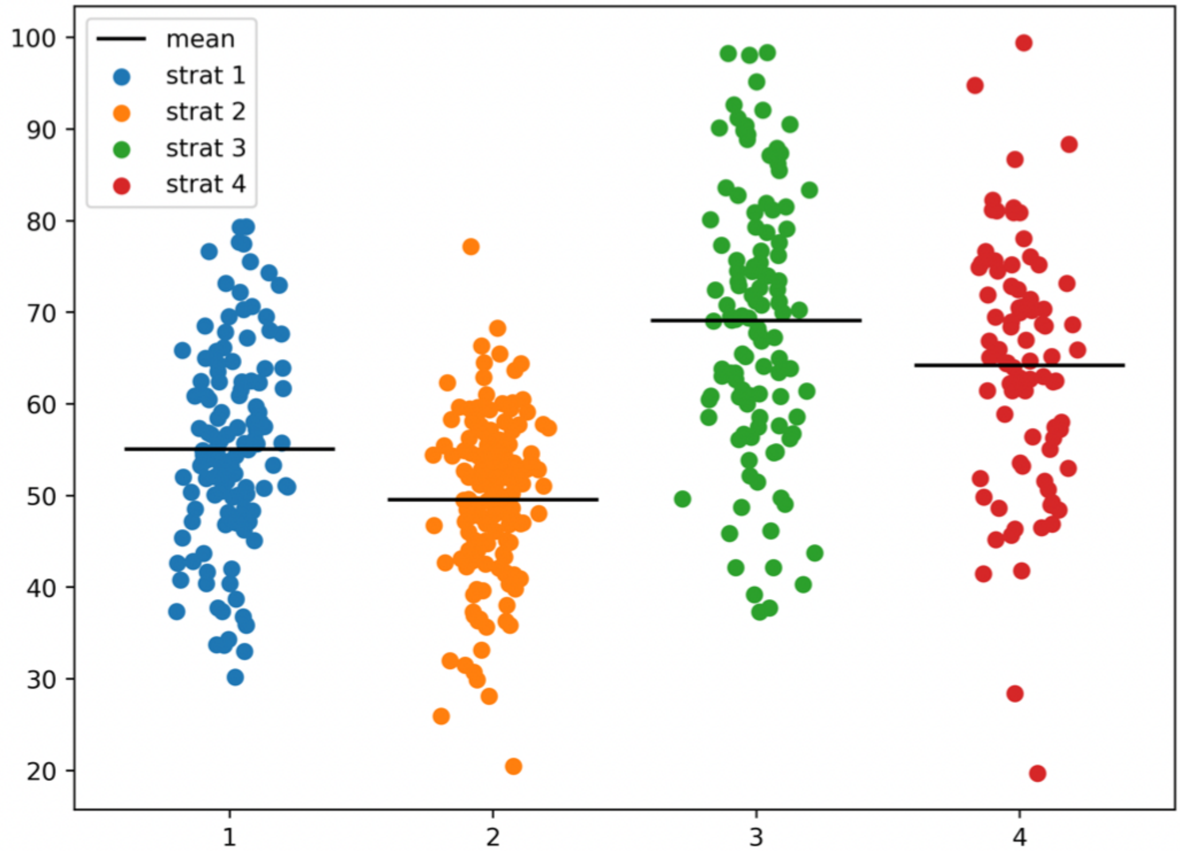
> Как проводить стратификацию?

Начнём с того, что такое стратификация.

Предположим, что нам удалось найти один или несколько признаков, которые коррелируют с исследуемой бизнес метрикой `Y`. Такие признаки `X` мы будем называть ковариатами. Эти величины должны быть измеримы до эксперимента.

Например, это могут быть пол, возраст или иные характеристики пользователя. Для международных онлайн-платформ хорошим признаком будет страна проживания пользователя.

Ковариаты используются для того, чтобы разделить всю генеральную совокупность на `K` непересекающихся подмножеств, называемых стратами.



Замечание: распределения целевой метрики в различных стратах должны отличаться, иначе стратификация не имеет смысла.

> Популяционное среднее

Нам необходимо оценить популяционное среднее бизнес метрики `Y`.

Введём обозначения для лекции:

$\mu = \mathrm{E}Y$ — популяционное среднее.

$\sigma^2 = VY$ — популяционная дисперсия.

μ_k, σ_k^2 — среднее значение и дисперсия бизнес-метрики для k-й страты.

w_k — доля k-й страты в популяции.

n_k — число пользователей из k-й страты в рассматриваемой группе.

$n = \sum\limits_{k=1}^{K} n_k$ — общий размер группы.

$Y_{11}, ... Y_{1n_1}, ... Y_{K1}, ... Y_{Kn_K}$ — выборка из ГС, где Y_{kj} — метрика для j-го пользователя k-й страты.

> Точечные оценки

Для популяционного среднего можно рассмотреть две несмещённые точечные оценки:

1. Простое среднее

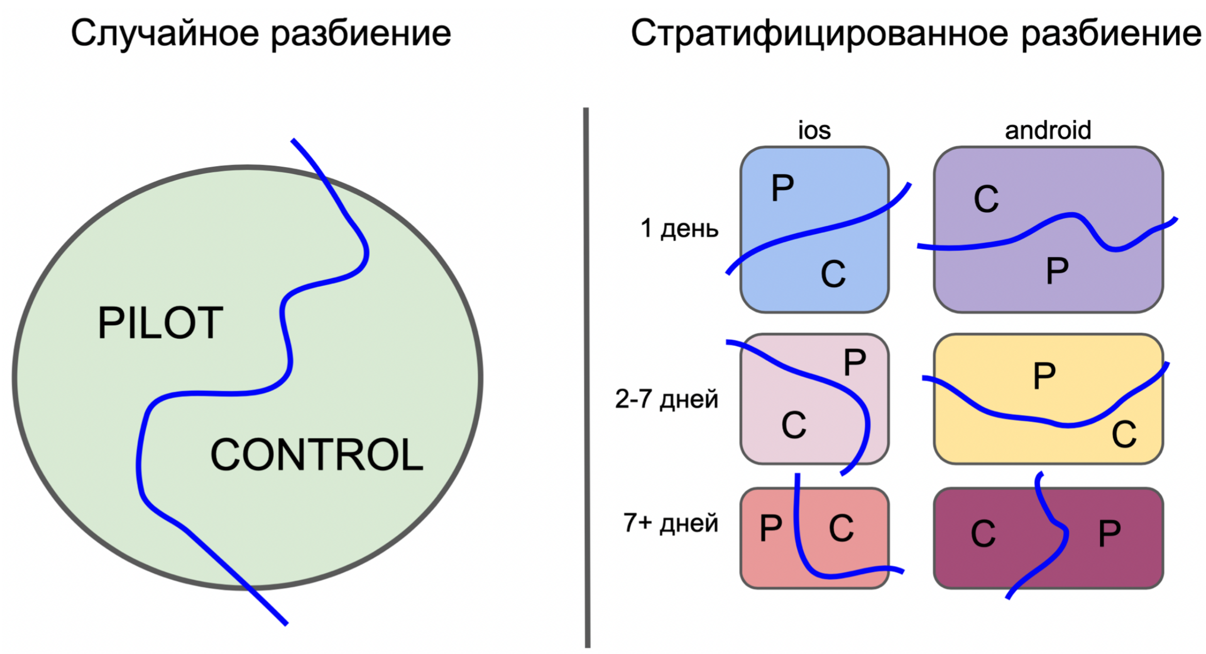
$$\overline{Y} = \frac{1}{n}\sum_{k=1}^{K}\sum_{j=1}^{n_k} Y_{kj}$$

2. Взвешенное среднее (стратифицированное среднее)

$$\hat{Y}_{strat} = \sum_{k=1}^{K} w_k \overline{Y}_k, \overline{Y}_k = \frac{1}{n_k}\sum^{n_k} j = 1Ykj$$

> Реализация

При стратифицированном разбиении необходимо контролировать, чтобы в каждой группе на протяжении всего эксперимента сохранялся баланс между контрольной и пилотной группами. В случайном разбиении этого может не происходить.



> Стратифицированное семплирование

Стратифицированное семплирование — метод понижения дисперсии. Для выборки мы должны обеспечить такие же доли каждой страты, что и в генеральной совокупности. Размер страт равен $n_k = nw_k$

Взвешенное среднее обычно используется для оценки популяционного среднего μ.

Способы семплирования:

1. Случайное семплирование — мы выбираем элементы без дополнительных требований к доле каждой из страт. Среднее и дисперсию для этого способа обозначим E_{srs} и V_{srs}
2. Стратифицированное семплирование — частота каждой страты должна быть такой же, как и в генеральной совокупности. Обозначения статистик E_{strat} и V_{strat}

Покажем, что в условиях стратифицированного семплирования две приведённые точечные оценки совпадают:

$$\hat{Y}_{strat} = \sum_{k=1}^{K} w_k \overline{Y}_k = \sum_{k=1}^{K} w_k \frac{1}{n_k} \sum^{n_k} j = 1Ykj = \sum_{k=1}^{K} \frac{n_k}{n} \frac{1}{n_k} \sum^{n_k} j = 1Ykj = \frac{1}{n} \sum_{k=1}^{K} \sum^{n_k} j = 1Ykj = \overline{Y}$$

Покажем, что эти точечные оценки — несмещённые оценки математического ожидания:

Случайное семплирование:

$$\mathrm{E}_{srs}(\overline{Y}) = \mathrm{E}_{srs}(\frac{1}{n} \sum_{k=1}^{K} \sum^{n_k} j = 1Ykj) = \frac{1}{n} \sum_{k=1}^{K} \sum^{n_k} j = 1\mathrm{E}srs(Y_{kj}) = \frac{1}{n} \sum_{k=1}^{K} \sum_{j=1}^{n_k} \mu = \mu$$

Стратифицированное семплирование:

$$\mathrm{E}_{strat}(\hat{Y}_{strat}) = \sum_{k=1}^{K} w_k \mathrm{E}_{strat}(\overline{Y}_k) = \sum_{k=1}^{K} w_k \mu_k = \mu$$

> Условное математическое ожидание

Условным математическим ожиданием измеримой функции $X : \Omega \to R$ относительно сигма-алгебры $\mathcal{G} \subseteq F$ называется $\mathcal{G}$-измеримая функция $\mathrm{E}(X|\mathcal{G}) : \Omega \to R$, такая, что для любого $A \in \mathcal{G}$ выполняется равенство:

$$\int_A X dP = \int_A \mathrm{E}(X|\mathcal{G}) dP$$

Пусть X и Y — случайные величины. $\mathrm{E}X < \infty$. Тогда условным математическим ожиданием случайной величины X относительно случайной величины Y назовём:

$$\mathrm{E}(X|Y) = \mathrm{E}(X|\sigma(Y)), \sigma(Y) = (Y^{-1}(B), B \in \mathrm{B})$$

> Полные математическое ожидание и дисперсия

> Закон полного математического ожидания

Найдём $\mathrm{E}(V(X|Y))$ и $V(\mathrm{E}(X|Y))$ для X и Y:

$$\mathrm{E}(V(X|Y)) = \mathrm{E}[\mathrm{E}[X^2|Y] - (\mathrm{E}[X|Y])^2] = \mathrm{E}[\mathrm{E}[X^2|Y]] - \mathrm{E}[(\mathrm{E}[X|Y])^2] = \mathrm{E}[X^2] - \mathrm{E}[(\mathrm{E}[X|Y])^2]$$

$$V(\mathrm{E}[X|Y]) = \mathrm{E}[(\mathrm{E}[X|Y])^2] - (\mathrm{E}[\mathrm{E}[X|Y]])^2 = \mathrm{E}[(\mathrm{E}[X|Y])^2] - (\mathrm{E}[X])^2$$

Из полученных равенств:

$$V(X) = \mathrm{E}[X^2] - (\mathrm{E}[X])^2 = \mathrm{E}[V(X|Y)] + V(\mathrm{E}[X|Y])$$

> Понижение дисперсии

> Межгрупповая и внутригрупповая дисперсия

Дисперсия случайного семплирования может быть представлена в виде суммы дисперсии внутри стратифицированной группы и между стратифицированными группами.

$$V_{srs}(Y) = E_{srs}(V_{srs}(Y|Z)) + V_{srs}(E_{srs}(Y|Z)) = E_{srs} \sum_{k=1}^{K} \sigma_k^2 l(Z = k) + V_{srs} \sum_{k=1}^{K} \mu_k l(Z = k) = \sum_{k=1}^{K} \sigma_k^2 E_{srs}(l(Z = k)) + E_{srs}(\sum_{k=1}^{K} \mu_k l(Z = k))^2 -$$

$$(E_{srs}\sum_{k=1}^{K}\mu_k l(Z=k))^2=\sum_{k=1}^{K}\sigma_k^2 w_k+\sum_{k=1}^{K}\mu_k^2 w_k-\mu^2=\sum_{k=1}^{K}\sigma_k^2 w_k+\sum_{k=1}^{K}w_k(\mu_k-mu)^2$$

> Дисперсия случайного семплирования

$$V_{srs}(\overline{Y})=\frac{1}{n}\sigma^2=\frac{1}{n}\sum_{k=1}^{K}w_k\sigma_k^2+\frac{1}{n}\sum_{k=1}^{K}w_k(\mu_k-\mu)2$$

> Дисперсия стратифицированного семплирования

$$V_{strat}(\hat{Y}_{strat})=\frac{1}{n}\sum_{k=1}^{K}w_k\sigma_k^2$$

> Понижение дисперсии

$$V_{srs}(\overline{Y})-V_{strat}(\hat{Y}_{strat})=\frac{1}{n}\sum_{k=1}^{K}w_k(\mu_k-\mu)^2$$

> Преимущества стратифицированного семплирования

- Стратифицированное среднее даёт несмещённую оценку популяционного среднего: $E_{strat}(\hat{Y}_{strat})=\mu$
- У этой оценки дисперсия ниже, чем при случайном семплировании:

  $$V_{srs}(\overline{Y})-V_{strat}(\hat{Y}_{strat})=\frac{1}{n}\sum_{k=1}^{K}w_k(\mu_k-\mu)^2$$

- В A/B-тестах мы можем получать большую чувствительность за счёт сниженной дисперсии.
- Подсчёт статистики по стратам и семплирование можно проводить прямо во время эксперимента. Например, выделяя каждого 100-го представителя страты и распределяя их между пилотом и контролем.

Но что делать, если семплирование не было произведено заранее?

> Постстратификация

> Проблемы случайного разбиения

Ситуация: имеем 200 000 пользователей для проведения пилота, среди них 100 пользователей старше 35 лет и пользуются iOS.

Какова вероятность, что отличие будет более чем в полтора раза, т.е. в одной из групп будет менее 40 пользователей старше 35 лет с iOS?

Количество пользователей в пилотной группе: $N_{pilot} \sim B(n = 100, p = 0.5)$

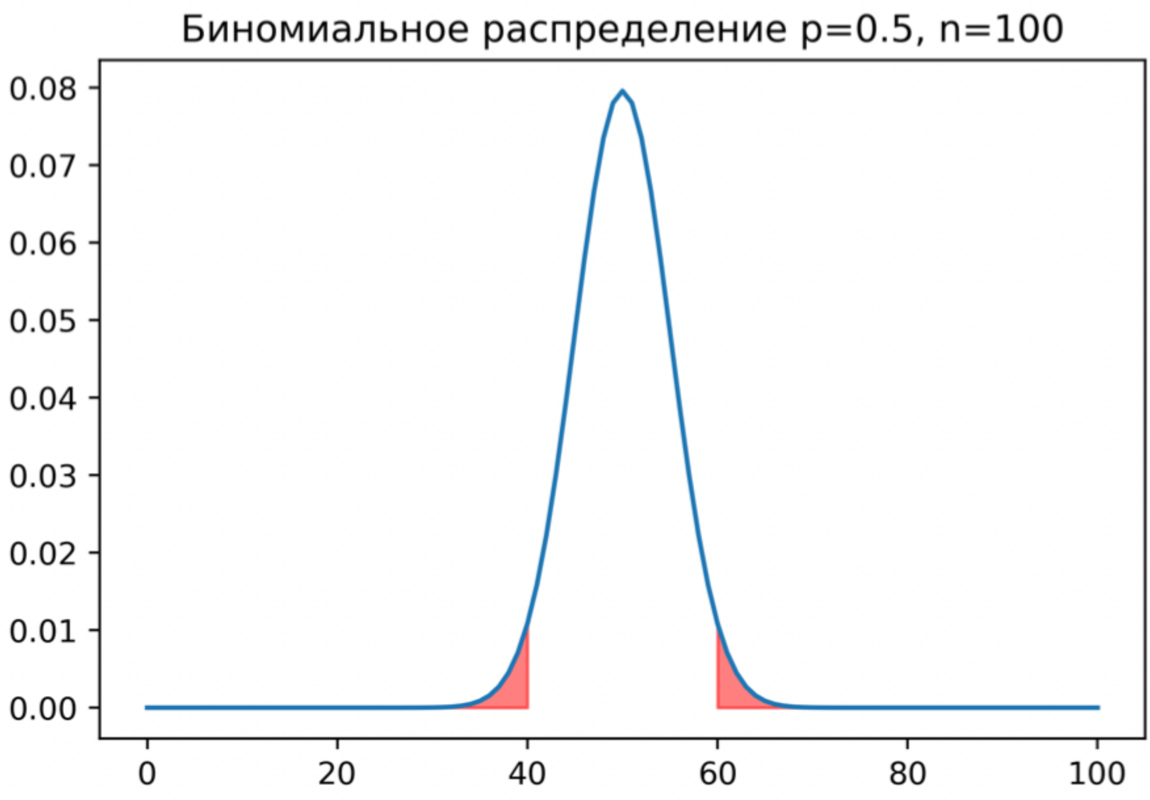
$P(\{N_{pilot} < 40\} \cup \{N_{control} < 40\}) \approx 0.06$

Вероятность перекоса более чем в полтора раза при различных N:

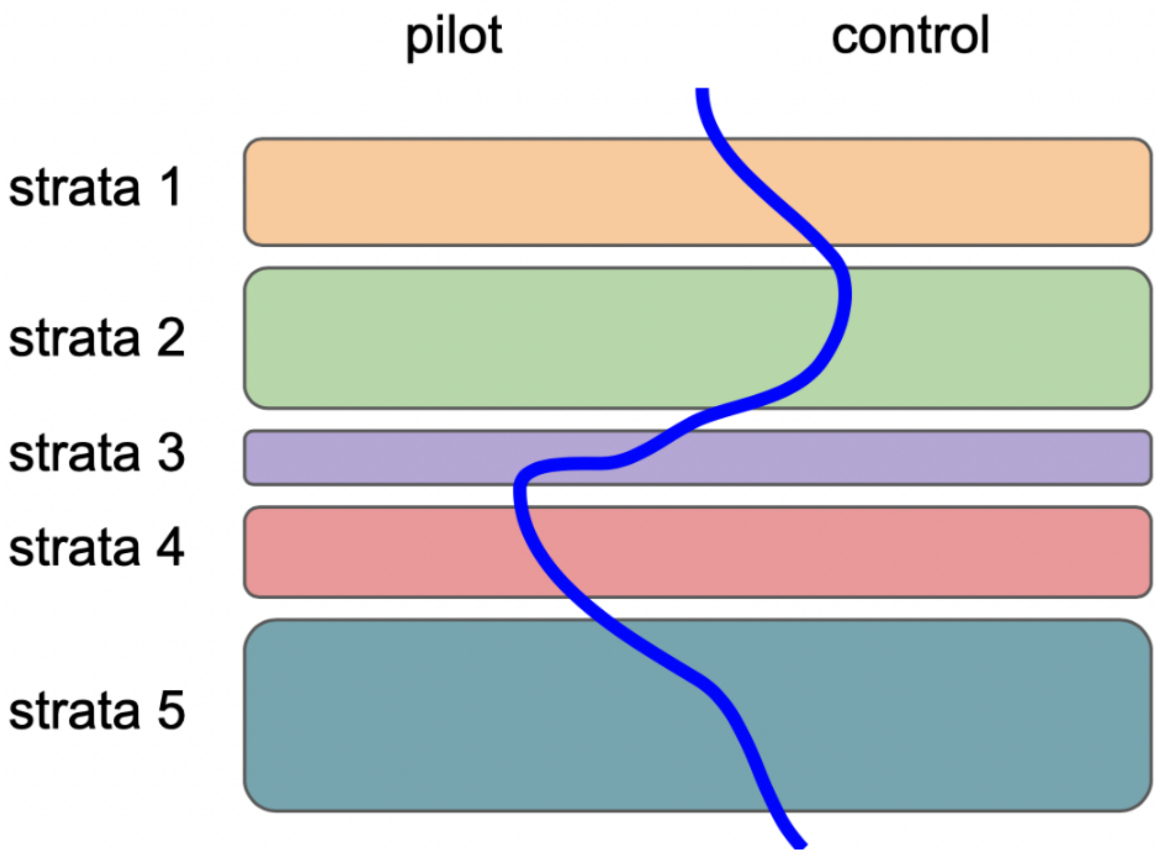
$N = 500, q \approx 0.001\%$

$N = 100, q \approx 6\%$

$N = 20, q \approx 50\%$



Что делать, если провели эксперимент без стратификации?



То есть в различных стратах группы pilot и control представлены в различных долях

Мы также можем заменить случайное среднее на стратифицированное среднее:

$$\hat{Y}_{strat} = \sum_{k=1}^{K} w_k \overline{Y}_k$$

Это соответствует перевзвешиванию каждой страты в соответствии с долей в генеральной совокупности.

Мы надеемся, что так можно снизить дисперсию.

> Дисперсия при постстратификации

Оценим дисперсию стратифицированного среднего при случайном семплировании:

$$V_{srs}(\hat{Y}_{strat}) = E_{srs}(V_{srs}(\hat{Y}_{strat}|n_1, ..., n_K)) +$$

$$V_{srs}(E_{srs}(\hat{Y}_{strat}|n_1, ..., n_K)) = E_{srs}(\sum_{k=1}^{K} w_k^2 V_{srs}(\overline{Y}_k|n_k)) +$$

$$V_{srs}(\sum_{k=1}^{K} w_k\mu_k) = E_{srs}(\sum_{k=1}^{K} w_k^2 \frac{1}{n_k}\sigma_k^2) + V_{srs}(\mu) = \sum_{k=1}^{K} w_k^2\sigma_k^2 E_{srs}(\frac{1}{n_k})$$

n_k — биномиальная случайная величина.

Дисперсия биномиальной случайной величины $V(n_k) = nw_k(1-w_k)$

Применим разложение Тейлора для функции $\frac{1}{n_k}$ в точке $\frac{1}{nw_k}$:

$$E_{srs}(\frac{1}{n_k}) = E_{srs}(\frac{1}{nw_k} - \frac{1}{n^2w_k^2}(n_k - nw_k) + \frac{1}{n^3w_k^3}(n_k - nw_k)^2) + O(\frac{1}{n^2}) = \frac{1}{nw_k} + \frac{1}{n^3w_k^3}E(n_k - nw_k)^2 + O(\frac{1}{n^2}) = \frac{1}{nw_k} + \frac{1}{n^3w_k^3}nw_k(n_k - nw_k) + O(\frac{1}{n^2}) = \frac{1}{nw_k} + \frac{1}{n^2w_k^2}(n_k - nw_k) + O(\frac{1}{n^2})$$

Получаем:

$$V_{srs}(\hat{Y}_{strat}) = \frac{1}{n}\sum_{k=1}^{K} w_k\sigma_k^2 + \frac{1}{n^2}\sum_{k=1}^{K}(1 - w_k)\sigma_k^2 + O(\frac{1}{n^2})$$

> Сравнение методов семплирования

> Дисперсии для различных методов

$$V_{srs}(\overline{Y}) = \frac{1}{n}\sum_{k=1}^{K} w_k\sigma_k^2 + \frac{1}{n}\sum_{k=1}^{K} w_k(\mu_k - \mu)^2$$

$$V_{strat}(\hat{Y}_{strat}) = \frac{1}{n}\sum_{k=1}^{K} w_k\sigma_k^2$$

$$V_{srs}(\hat{Y}_{strat}) = \frac{1}{n}\sum_{k=1}^{K} w_k\sigma_k^2 + \frac{1}{n_2}\sum_{k=1}^{K}(1 - w_k)\sigma_k^2 + O(\frac{1}{n^2})$$

> Соотношения дисперсий

$$V_{strat}(\hat{Y}_{strat}) = V_{srs}(\hat{Y}_{strat}) + O(\frac{1}{n^2}) = V_{srs}(\overline{Y}) + O(\frac{1}{n})$$

$$V_{strat}(\hat{Y}_{strat}) \le V_{srs}(\hat{Y}_{strat}) \le V_{srs}(\overline{Y})$$

> Оценка пилота

С помощью бутстрепа оцениваем распределение разности средних стратифицированных $\hat{Y}_{strat}$:

- Семплируем данные пилотной и контрольной групп;
- Считаем разность стратифицированных средних: $\hat{Y}^{bs}_{strat} - \hat{X}^{bs}_{strat}$
- Строим доверительный интервал;
- Проверяем, входит ли ноль в доверительный интервал.

С помощью теста Стьюдента:

- Считаем стратифицированные средние: $\hat{Y}_{strat}$
- Считаем оценку дисперсий: $\sigma^2_Y = V(\hat{Y}_{strat}) \approx \frac{1}{n}\sum_{k=1}^{K} w_k \sigma_k^2$
- Считаем t-статистику и p-value:

$$t = \frac{\hat{Y}_{strat} - \hat{X}_{strat}}{\sqrt{\frac{\sigma_Y^2 + \sigma_X^2}{n}}}$$

> Резюме

1. Познакомились с методом стратификационного семплирования.
2. Убедились, что этот метод значительно снижает дисперсию и его полезно использовать в A/B-экспериментах.
3. Если забыли провести стратификацию, то на выручку может прийти постстратификация.

> Материалы для самостоятельного изучения

1. W. G. Cochran. Sampling Techniques. Wiley, 1977
2. Improving the Sensitivity of Online Controlled Experiments: Case Studies at Netflix